响应变量删失情况下线性模型的 FIC 模型选择和模型平均

孙志猛*, 马景义, 苏治

中央财经大学统计学院, 北京 100081

E-mail: sunzhimeng99@126.com, jingyima@yahoo.com.cn, 15801237605@163.com

收稿日期: 2012-09-14; 接受日期: 2013-05-03; * 通信作者

中央财经大学第二批青年科研创新团队、2012 全国统计科学研究(计划)、中央财经大学学科建设经费、2012 年北京市社科联青年社科人才、国家自然科学基金(批准号: 71101157)和教育部人文社会科学研究青年基金(10YJC790220)资助项目

摘要 本文给出了响应变量随机右删失情况下线性模型的 FIC (focused information criterion) 模型选择方法和光滑 FIC 模型平均估计方法, 证明了兴趣参数的 FIC 模型选择估计和光滑 FIC 模型平均估计的渐近正态性, 通过随机模拟研究了估计的有限样本性质, 模拟结果显示, 从均方误差和一定置信水平置信区间的经验覆盖概率看, 兴趣参数的光滑 FIC 模型平均估计均优于 FIC, AIC (Akaike information criterion) 和 BIC (Bayesian information criterion) 等模型选择估计; 而 FIC 模型选择估计与 AIC 和 BIC 等模型选择估计相比, 也表现出了一定的优越性. 通过分析原发性胆汁性肝硬化数据集, 说明了本文方法在实际问题中的应用.

关键词 随机右删失 模型选择 模型平均 估计理论

MSC (2010) 主题分类 62N01, 62G05

1 引言

在现代统计实践中, 随着科技的发展和数据采集工具的进步, 研究者往往能够测量到更多的变量. 然而, 在回归建模初期, 确定把哪些变量引入模型是整个建模过程非常重要的一步. 这类问题称作变量选择或模型选择. 20 世纪 70 年代以来, 模型选择问题一直是统计研究的热点课题, 涌现了许多著名的变量选择准则和研究成果, 如 AIC 准则^[1]、BIC 准则^[2]、 C_p 准则^[3]、LASSO (least absolute shrinkage and selection operator) 准则^[4] 和 SCAD (smoothly clipped absolute deviation) 准则^[5] 等. 这些传统的模型选择方法致力于选择出某个“好”的模型, 希望该模型在数据分析的各个方面都具有好的性质, 以用于预测和解释数据信息的各个方面. 然而, Claeskens 和 Hjort^[6] 的数据分析实例表明, 当研究者的兴趣是估计某个特定的参数时, 如模型参数的某个特定函数, 这些传统的模型选择方法选出的模型并不一定是“好”的模型. 事实上, 对一些参数而言, 某个备选模型可能是“好”的, 但对另外一些参数而言, 另外的某个备选模型可能更优. Claeskens 等人^[7] 对 Wisconsin 糖尿病性视网膜病变流行病学研究的数据分析也印证了这一点. 在该问题的驱动下, Claeskens 和 Hjort^[6] 提出了一个新的模型选择准则, 即 FIC 准则. FIC 本质上是兴趣参数估计极限风险的无偏估计, 其允许并且鼓励对不同的兴趣参数选取不同的备选模型. 近年来, FIC 准则被应用于许多模型的模型选择, 如 Hjort 和 Claeskens^[8]

英文引用格式: Sun Z M, Ma J Y, Su Z. FIC model selection and model averaging for linear model with censored response (in Chinese). *Sci Sin Math*, 2013, 43: 647–661, doi: 10.1360/012012-456

给出了比例风险模型的 FIC 模型选择方法; Zhang 和 Liang^[9] 给出了广义可加部分线性模型的 FIC 模型选择方法; Du 等人^[10] 提出了基于分位数回归的 FIC 模型选择. 这些模型选择方法在一定的条件下解决了引入哪些解释变量才能使所选模型相对较“优”的问题. 然而, 越来越多的研究成果表明, 这些方法也有着本身固有的不足. Yuan 和 Yang^[11] 指出了这种模型选择方法的稳健性不够理想, 数据较小的波动, 往往有可能导致不同的模型选择结果; Leeb 和 Pötscher^[12, 13] 指出了模型选择方法会忽略模型选择阶段存在的随机性, 从而可能给统计推断带来严重后果, 例如得到的置信水平为 95% 的置信区间的实际置信度可能仅为 75%, 甚至更低, 又或者一定检验水平的检验方法的实际水平可能会更高等; Leung 和 Barron^[14] 指出了模型选择方法无法避免选到“差的模型”的风险. 为了克服这些不足, 一种有效方法是对不同模型下的估计量进行加权平均. 这种方法称作模型平均法. 由于考虑了不同模型下的数据信息, 模型平均方法可以防止模型选择的不确定性, 避免选择单一模型潜在的风险, 减小估计的均方误差, 提高覆盖概率 (参见文献 [14, 15]), 稳健性也更强. 对模型平均方法的研究最早可追溯到 Bates 和 Granger^[16]. 近年来, 模型平均方法成为统计学国际前沿课题, 涌现了大量的研究成果. 其中一个重要的分支是频率模型平均方法, 如 Buckland 等人^[17] 基于 AIC 准则和 BIC 准则讨论了模型平均权重的光滑 AIC 和光滑 BIC 方法; Hansen^[18] 基于 Mallows 准则讨论了模型平均估计的权重选择方法; Hjort 和 Claeskens^[19] 提出了模型平均的光滑 FIC 方法; Liang 等人^[20] 提出了模型平均的最优 (optimal) 权重选择方法; Deng 和 Liang^[21] 研究了半参数可加部分线性模型的模型平均方法; Zhang 和 Liang^[9] 研究了广义可加部分线性模型的模型平均, 并给出了模型平均估计的渐近正态性; Du 等人^[10] 把分位数回归和模型平均技术结合起来, 给出了基于分位数回归的模型平均方法, 有效地解决了随机误差分布是厚尾的情况; Sun^[22] 考虑了因子线性模型的光滑 FIC 模型平均估计. 这些研究成果, 对回归建模理论作出了重要推进. 然而, 这些研究基本是对独立同分布样本进行的, 而对于现代统计实践中经常出现的删失数据, 还不能有效解决. 鉴于删失数据在现代统计实践, 尤其是临床医学和保险领域的广泛存在性, 在这些工作的基础上, 对删失数据的模型平均方法的研究就显得尤为迫切. 另一方面, 删失数据本身结构的复杂性使得其理论研究也具有较大难度. 这是因为对于删失数据而言, 由于删失变量的存在会带来数据的不平衡性, 一般需要在参数的估计方程中调整删失变量带来的偏差, 然而这一调整往往会使估计方程不再是独立随机变量的和, 从而给参数估计渐近正态性的证明带来较大难度. 目前, 对于响应变量被随机删失情况下的回归模型的模型平均估计的研究还非常少见, 本文拟对这一问题进行探索.

删失数据是现代统计实践中常见的复杂数据类型. 近年来, 对删失数据统计分析方法的研究是一个国际热点课题. Buckley^[23], Koul 等人^[24] 和 Leurgans^[25] 通过构造不同的无偏变换变量, 研究了删失数据下线性模型的估计问题; Wang 和 Li^[26] 用经验似然方法研究了随机右删失数据下部分线性模型的统计推断理论; Stute^[27, 28] 提出并证明了响应变量删失情况下线性模型的加权最小二乘估计的相合性和渐近正态性; Li 和 Wang^[29] 考虑了删失变量随机缺失情况下的线性模型的加权最小二乘估计; Bang 和 Tsiatis^[30] 研究了删失数据下线性模型的加权中位数回归; Peng 和 Huang^[31] 研究了删失数据的加权分位数回归问题; Shows 等人^[32] 研究了删失分位数回归的模型选择问题; Sun 等人^[33] 考虑了响应变量删失和协变量带有测量误差情况下部分线性单调回归模型的估计. 这些研究成果丰富了删失数据的理论研究体系. 然而, 这些成果往往是在固定模型或模型选择的框架下进行的, 而对于模型平均问题的研究还没有看到. 而删失数据结构的复杂性也使得对于这一问题的研究具有一定难度. 本文对这一问题进行研究, 给出删失线性模型的 FIC 模型选择方法和基于 FIC 的模型平均方法.

本文剩余的部分组织结构如下: 第 2 节给出每个子模型下参数的估计及其渐近性质; 第 3 节给出 FIC 模型选择和模型平均方法; 第 4 节对本文所提估计方法的有限样本性质进行 Monte Carlo 模拟研

究; 第 5 节给出一个实际例子说明所提方法在实际问题中的应用; 第 6 节概括本文的结论, 并讨论将来的研究计划; 第 7 节给出本文主要结果的证明.

2 子模型下参数的估计及其渐近性质

考虑如下的线性模型:

$$T = X^T \beta_0 + \varepsilon, \quad (2.1)$$

其中, T 为响应变量, $X = (X_1, \dots, X_d)^T$ 为 d 维解释变量, ε 为随机误差, β_0 为 d 维未知回归参数. 在许多实际问题中, 响应变量 T 可能为个体的生存时间或者生存时间的某个变换. 在许多情况下, 尤其是在生物医学和保险学研究领域, T 往往是不能完全观测的, 一种常见的情形是 T 被某个删失时间变量随机右删失. 例如, 某些个体中途单方面退出研究, 或者某些个体在试验结束时仍然存活, 因此, 观测不到生存时间. 假设删失时间变量为 C , 此时, 研究者仅能观测到 $Y = \min(T, C)$. 假设模型 (2.1) 的独立同分布样本为 $\{(Y_i, X_i, \delta_i), i = 1, \dots, n\}$, 其中, $\delta_i = I(T_i \leq C_i)$, $I(\cdot)$ 为示性函数. 本文假设 C 与 T 独立, T 和 C 的概率分布函数 F 和 G 均为连续函数.

不失一般性, 假设回归参数 β_0 可以分为两个部分 $\beta_0 = (\check{\beta}^T, \bar{\beta}^T)^T$. 其中, 第一个部分 $\check{\beta}$ 包含了 β_0 确定进入模型的分量, 假设为前 $\check{d}$ 个; 第二个部分 $\bar{\beta}$ 包含了 β_0 的剩余 $\bar{d} = d - \check{d}$ 个待引入模型的参数分量. 与文献 [19, 21] 一样, 为了得到模型平均估计的大样本性质, 我们假定 β_0 具有局部结构,

$$\beta_0 = \left(\check{\beta}_0^T, \frac{\eta^T}{\sqrt{n}} \right)^T, \quad \eta = (\eta_1^T, \dots, \eta_{\bar{d}}^T)^T$$

为 $\bar{d}$ 维常向量.

为了得到兴趣参数的模型平均估计, 需要对每个子模型下兴趣参数的估计进行加权平均. 因此, 首先需要定义每一个子模型下参数的估计. 为此, 记 $\beta_S = (\check{\beta}_S^T, \bar{\beta}_S^T)^T$ 为第 S 个子模型的回归参数, 令 π_S 为把 $\bar{\beta}$ 映射到 $\bar{\beta}_S$ 的映射矩阵, 其维数为 $\bar{d}_S \times \bar{d}$, $\bar{\beta}_S$ 为第 S 个子模型包含的 $\bar{\beta}$ 的分量构成的向量. 故第 S 个子模型的解释变量为 $\Pi_S X$, 其中, $\Pi_S = \text{diag}(I_{\check{d}}, \pi_S)$, $I_{\check{d}}$ 是 $\check{d}$ 维单位阵.

对于完全数据, 参数 β_0 的估计可以通过最小化目标函数

$$L_n(\beta) = \sum_{i=1}^n (Y_i - X_i^T \beta)^2 \quad (2.2)$$

得到, 或者等价地通过求解估计方程

$$n^{-1} \sum_{i=1}^n X_i (Y_i - X_i^T \beta) = 0$$

得到. 由于删失变量的存在导致了数据的不平衡性, 因此, 为了在删失数据下得到 β_0 的估计方程, 可以考虑对 (2.2) 进行调整, 以纠正删失变量带来的偏倚. 注意到第 i 个体在 T_i 不被删失的概率为

$$P(C_i \geq T_i) = 1 - G(T_i),$$

因此, 平均而言, 每个未被删失的个体实际上代表了 $1/(1 - G(T_i))$ 个个体. 事实上, 容易证明

$$E[\delta_i / (1 - G(T_i))] = 1.$$

因此, 可以用未被删失观测个体的生存时间定义如下加权目标函数:

$$L_n(\beta_S) = \sum_{i=1}^n \frac{\delta_i}{1 - G(Y_i)} (Y_i - X_i^T \beta)^2. \quad (2.3)$$

目标函数 (2.3) 中包含有未知函数 G , 因此还需要对 G 进行估计. 一个自然的想法是用其 Kaplan-Meier 估计^[34] 进行替代, 其定义为

$$\hat{G}(y) = 1 - \prod_{i=1}^n \left[\frac{n-i}{n-i+1} \right]^{I[Y_{(i)} \leq y, \delta_{(i)}=0]},$$

其中, $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$ 为 Y_i 的顺序统计量, $\delta_{(i)}$ 为对应于 $Y_{(i)}$ 的示性变量. 由此得到估计的加权目标函数

$$L_n(\beta_S) = \sum_{i=1}^n \frac{\delta_i}{1 - \hat{G}(Y_i)} (Y_i - X_i^T \beta)^2. \quad (2.4)$$

易见, 目标函数 (2.4) 的极小化问题等价于求解估计方程

$$S_n(\beta) = \sum_{i=1}^n \frac{\delta_i}{1 - \hat{G}(Y_i)} X_i (Y_i - X_i^T \beta) = 0.$$

在第 S 个子模型, 为了得到 β_S 的估计 $\hat{\beta}_S$, 定义如下目标函数:

$$L_n(\beta_S) = \sum_{i=1}^n \frac{\delta_i}{1 - \hat{G}(Y_i)} (Y_i - (\Pi_S X_i)^T \beta_S)^2.$$

类似地, $\hat{\beta}_S$ 可以通过求解估计方程

$$S_n(\beta_S) = \sum_{i=1}^n \frac{\delta_i}{1 - \hat{G}(Y_i)} \Pi_S X_i (Y_i - (\Pi_S X_i)^T \beta_S) = 0$$

得到. 不难求出

$$\hat{\beta}_S = \left[\sum_{i=1}^n \frac{\delta_i}{1 - \hat{G}(Y_i)} \Pi_S X_i X_i^T \Pi_S^T \right]^{-1} \sum_{i=1}^n \frac{\delta_i}{1 - \hat{G}(Y_i)} \Pi_S X_i Y_i.$$

下面的定理 1 给出了每一个子模型下 $\hat{\beta}_S$ 的渐近正态性.

定理 1 假设如下条件成立,

(A₁) 矩阵 $\Sigma = E[XX^T]$, $\Omega_1 = E[\frac{XX^T \varepsilon^2}{1-G(Y)}]$, $E[XX^T \varepsilon^2]$ 均为有限正定阵;

(A₂) $\sup_x E[\frac{Y^2}{(1-G(Y))^3} | X = x] < \infty$, $E[\frac{\|X\|Y}{((1-G(Y))(1-F(Y)))^{\frac{1}{2}}}] < \infty$;

(A₃) 记 $\tau_G = \inf\{y : G(y) = 1\}$, $\tau_F = \inf\{y : F(y) = 1\}$, $\tau_F < \tau_G$, 令 Y 的分布函数为

$$Q(y) = 1 - (1 - F(y))(1 - G(y)), \quad \tau_Q = \inf\{y : Q(y) = 1\};$$

(A₄) $\int_0^{\tau_Q} \|B(y)\|^2 (1 - \Lambda_G(y)) dG(y) < \infty$, 其中,

$$B(y) = \frac{E[X \varepsilon I[Y > y]]}{(1 - G(y))(1 - F(y))}, \quad \Lambda_G(y) = \int_0^y (1 - G(s))^{-1} dG(s);$$

(A₅) 随机向量 $\phi = X\varepsilon$ 以概率有界,
则有

$$\sqrt{n}(\Pi_S^T \hat{\beta}_S - \beta_0) \xrightarrow{d} \Pi_S^T (\Pi_S \Sigma \Pi_S^T)^{-1} \Pi_S W + A_S \begin{pmatrix} 0 \\ \eta \end{pmatrix}, \quad n \rightarrow \infty,$$

其中,

$$W \sim N(0, \Omega), \quad \Omega = \Omega_1 + \Omega_2 + 2\Omega_3, \quad \Omega_2 = \int_0^{\tau_Q} (B(y))^{\otimes 2} (1 - F(y)) dG(y),$$

$$\Omega_3 = - \int_0^{\tau_Q} E\{X\varepsilon B^T(y) I[Y \geq y]\} d\Lambda_G(y), \quad A_S = (\Pi_S^T (\Pi_S \Sigma \Pi_S^T)^{-1} \Pi_S \Sigma - I)(I - \Pi_S^T \Pi_S),$$

“ $\xrightarrow{d}$ ” 表示以分布收敛.

3 FIC 和模型平均

假设感兴趣的参数为 β_0 的某个已知函数 $\phi_0 = \phi(\beta_0)$. 对于第 S 个子模型, 我们可以用

$$\hat{\phi}_S = \phi(\Pi_S^T \hat{\beta}_S)$$

来估计 ϕ_0 . 下面定理 2 给出 $\hat{\phi}_S$ 的渐近正态性. 在给出定理之前, 假设以下条件成立,

(A₆) ϕ 在 β_0 有如下意义可微: 存在一个线性变换 $\phi'_{\beta_0}: \mathbb{R}^d \mapsto \mathbb{R}$ 使得

$$\phi(\beta_0 + h) - \phi(\beta_0) = \phi'_{\beta_0} h + o(\|h\|), \quad h \rightarrow 0,$$

其中, $\|h\|$ 为 Euclid 范数. 易见, 若 ϕ 可微, 则其偏导数存在, 且其导数变换 $h \rightarrow \phi'_{\beta_0} h$, 由

$$\phi'_{\beta_0} = \left(\frac{\partial \phi_1}{\partial t_1}(\beta_0) \cdots \frac{\partial \phi_1}{\partial t_d}(\beta_0) \right)$$

定义.

定理 2 在假设条件 (A₁)–(A₆) 下, 有

$$\sqrt{n}(\hat{\phi}_S - \phi_0) \xrightarrow{d} \phi'_{\beta_0} \Pi_S^T (\Pi_S \Sigma \Pi_S^T)^{-1} \Pi_S W + \phi'_{\beta_0} A_S \begin{pmatrix} 0 \\ \eta \end{pmatrix} \triangleq \Gamma_S, \quad n \rightarrow \infty.$$

下面定义第 S 个子模型关于 ϕ_0 的 FIC 值. 令 $R_S = \Pi_S^T (\Pi_S \Sigma \Pi_S^T)^{-1} \Pi_S$. 基于定理 2, 可得用 $\sqrt{n}\hat{\phi}_S$ 估计 $\sqrt{n}\phi_0$ 的均方误差为

$$E(\Gamma_S \Gamma_S^T) = \phi'_{\beta_0} \left\{ R_S \Omega R_S + A_S \begin{pmatrix} 0 & 0 \\ 0 & \eta \eta^T \end{pmatrix} A_S^T \right\} \phi_{\beta_0}^T. \quad (3.1)$$

在模型选择时, 使 (3.1) 达到最小的模型可以认为是 “最优” 模型. 然而 (3.1) 为均方误差的理论形式, 根据样本做模型选择时需要对其中的未知量进行估计. 注意到 $E(\Gamma_S \Gamma_S^T)$ 中的 Σ 和 Ω 均易得到相合估计, 因此, 关键问题是估计未知参数 η . 令 $\hat{\beta}_{\text{full}}$ 为全模型下 β 的估计, 则 $\hat{\eta} = \sqrt{n}[0, I]\hat{\beta}_{\text{full}}$ 为全模型下 η 的估计. 在定理 1 中令 $\Pi_S = I_p$ 可知,

$$\hat{\eta} \xrightarrow{d} \Delta \equiv [0, I]\Sigma^{-1}W + \eta,$$

$$\Delta \sim N(\eta, [0, I]\Sigma^{-1}\Omega\Sigma^{-1}[0, I]^T).$$

因此, $\hat{\eta}$ 作为 η 的估计并不是相合的. 然而, 注意到,

$$E[\Delta\Delta^T] = \eta\eta^T + [0, I]\Sigma^{-1}\Omega\Sigma^{-1}[0, I]^T,$$

因此, 可用 $\hat{\eta}\hat{\eta}^T - [0, I]\Sigma^{-1}\Omega\Sigma^{-1}[0, I]^T$ 估计 (3.1) 中的 $\eta\eta^T$. 将其代入 (3.1) 得 $E(\Gamma_S\Gamma_S^T)$ 的估计, 即第 S 个子模型关于 ϕ_0 的 FIC 值,

$$\begin{aligned} \text{FIC}_S = & \phi'_{\beta_0} \left[R_S \Omega R_S + A_S \begin{pmatrix} 0 & 0 \\ 0 & \hat{\eta}\hat{\eta}^T \end{pmatrix} A_S^T \right] \phi'_{\beta_0} \\ & - \phi'_{\beta_0} A_S \begin{pmatrix} 0 & 0 \\ 0 & I_{\bar{d}} \end{pmatrix} \Sigma^{-1} \Omega \Sigma^{-1} \begin{pmatrix} 0 & 0 \\ 0 & I_{\bar{d}} \end{pmatrix} A_S^T \phi'_{\beta_0}. \end{aligned}$$

基于 FIC 值, 我们可以进行模型选择. 使得 FIC 值达到最小的子模型称为最优模型, 相应的参数估计称为 FIC 估计. 同时, 基于每个子模型的 FIC 值, 可以定义 ϕ_0 的光滑 FIC 模型平均估计如下:

$$\hat{\phi} = \sum_S w(S | \hat{\eta}) \hat{\phi}_S.$$

本文中权函数的选取与文献 [9] 和 [19, (5.4)] 类似, 取为

$$w(S | \hat{\eta}) = \exp \left\{ \frac{\text{FIC}_S}{\phi'_{\beta_0} \Sigma^{-1} \Omega \Sigma^{-1} \phi'_{\beta_0}} \right\} / \sum_S \exp \left\{ \frac{\text{FIC}_S}{\phi'_{\beta_0} \Sigma^{-1} \Omega \Sigma^{-1} \phi'_{\beta_0}} \right\},$$

其中, $\Sigma^{-1} \Omega \Sigma^{-1}$ 为常风险, 即用 $\hat{\eta}$ 作为 η 的估计时的方差.

下面的定理 3 给出了 $\hat{\phi}$ 的渐近性质.

定理 3 假设条件 (A₁)–(A₆) 成立, 则

$$\sqrt{n}(\hat{\phi} - \phi_0) \xrightarrow{d} Q_1(\Delta)W + Q_2(\Delta) \begin{pmatrix} 0 \\ \eta \end{pmatrix}, \quad n \rightarrow \infty,$$

其中,

$$Q_1(\cdot) = \sum_S w(S | \cdot) \phi'_{\beta_0} R_S, \quad Q_2(\cdot) = \sum_S w(S | \cdot) \phi'_{\beta_0} A_S.$$

记 $\hat{\Lambda} = Q_1(\Delta) \hat{\Omega} Q_1(\Delta)^T$, 由定理 3 知,

$$\hat{\Lambda}^{-\frac{1}{2}} \left[\sqrt{n}(\hat{\phi} - \phi_0) - Q_2(\Delta) \begin{pmatrix} 0 \\ \eta \end{pmatrix} \right] \xrightarrow{d} N(0, 1).$$

由此, 可以得到 ϕ_0 的置信度为 $1 - \alpha$ 的近似置信区间 $[\text{low}_n, \text{up}_n]$, 其中,

$$\text{low}_n = \hat{\phi} - Q_2(\Delta) \begin{pmatrix} 0 \\ \eta \end{pmatrix} / \sqrt{n} - z_{\frac{\alpha}{2}} \hat{\Lambda}^{-\frac{1}{2}} / \sqrt{n},$$

$$\text{up}_n = \hat{\phi} - Q_2(\Delta) \begin{pmatrix} 0 \\ \eta \end{pmatrix} / \sqrt{n} + z_{\frac{\alpha}{2}} \hat{\Lambda}^{-\frac{1}{2}} / \sqrt{n},$$

$z_{\frac{\alpha}{2}}$ 为标准正态分布的 $\frac{\alpha}{2}$ 分位数.

4 模拟研究

本节通过 Monte Carlo 随机模拟研究估计的有限样本性质. 在模拟时, 考虑如下线性模型:

$$T_i = X_i^T \beta_0 + \varepsilon_i, \quad i = 1, 2, \dots, n,$$

其中,

$$\beta_0 = \left(1, -2, \frac{2}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \frac{3}{\sqrt{n}}\right)^T, \quad X_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}) \sim N_5(0, \Sigma_X),$$

这里, 5 维协方差阵 Σ_X 的第 i 行第 j 列的元素为 $\Sigma_{Xij} = 0.5^{|i-j|}$, $1 \leq i, j \leq 5$, $\varepsilon_i \sim N(0, 1)$. 删失变量 C_i 服从区间 $[a, b]$ 上的均匀分布, 通过调整 a 和 b 来控制删失比例. 为了进行模型平均估计, 我们把 X_1 和 X_2 设定为每个子模型的必选变量, X_3, X_4 和 X_5 设定为每个子模型的备选变量进行模型平均. 因此, 共有 $2^3 = 8$ 个子模型可供模型选择和模型平均. 我们设定感兴趣的参数有 $\phi_1 = \beta_1$, $\phi_2 = \beta_2$, $\phi_3 = \beta_1 - \beta_2 + \beta_3$. 在模拟中, 分别考虑 20% 和 40% 两种删失比例. 在两种删失比例下, 样本量分别取 $n = 100, 200$ 和 300 , 共产生了 $M = 1000$ 个样本. 为了研究本文所提 FIC 模型选择方法和光滑模型平均方法 (SFIC) 在有限样本量下的表现, 我们计算了估计的均方误差 (MSE),

$$\text{MSE}(\hat{\phi}) = \frac{1}{M} \sum_{m=1}^M (\hat{\phi}_m - \phi)^2$$

和相应的置信度为 95% 的置信区间的经验覆盖概率 (CP), 其中, $\hat{\phi}_m$ 为第 m 个样本下 ϕ 的估计. 作为与传统模型选择方法的比较, 我们也计算了 AIC 准则和 BIC 准则下的模型选择的相关结果. 模拟结果见表 1 和 2. 从表 1 和 2 不难看出,

(1) 在固定的删失比例和样本量下, 总体来说, SFIC 所得估计的均方误差比 FIC, AIC 和 BIC 小; SFIC 所得置信区间的经验覆盖概率比 FIC, AIC 和 BIC 大. 因此, 从均方误差和经验覆盖概率两个角度看, SFIC 均比 FIC, AIC 和 BIC 的估计更为精确; FIC, AIC 和 BIC 三种模型选择方法的均方误差

表 1 删失概率约为 20% 时不同准则下参数估计的 MSE 和 CP

	准则	ϕ_1		ϕ_2		ϕ_3	
		MSE	CP	MSE	CP	MSE	CP
$n = 100$	SFIC	0.01690	0.934	0.02105	0.929	0.09760	0.920
	FIC	0.01725	0.933	0.02256	0.930	0.11418	0.920
	AIC	0.01673	0.938	0.02228	0.912	0.11004	0.839
	BIC	0.01703	0.934	0.02493	0.877	0.12874	0.785
$n = 200$	SFIC	0.00814	0.941	0.01106	0.928	0.05093	0.932
	FIC	0.00827	0.938	0.01223	0.926	0.06053	0.931
	AIC	0.00818	0.936	0.01169	0.903	0.05742	0.830
	BIC	0.00824	0.934	0.01342	0.870	0.06843	0.757
$n = 300$	SFIC	0.00527	0.950	0.00714	0.937	0.03303	0.939
	FIC	0.00536	0.945	0.00763	0.935	0.03839	0.941
	AIC	0.00530	0.945	0.00746	0.921	0.03662	0.840
	BIC	0.00534	0.947	0.00852	0.889	0.04487	0.769

表 2 删失概率约为 40% 时不同准则下参数估计的 MSE 和 CP

	准则	ϕ_1		ϕ_2		ϕ_3	
		MSE	CP	MSE	CP	MSE	CP
$n = 100$	SFIC	0.02653	0.920	0.03200	0.918	0.13811	0.915
	FIC	0.02709	0.915	0.03508	0.917	0.16415	0.912
	AIC	0.02648	0.907	0.03387	0.879	0.15563	0.827
	BIC	0.02693	0.906	0.03678	0.863	0.17312	0.785
$n = 200$	SFIC	0.01119	0.938	0.01612	0.927	0.07068	0.920
	FIC	0.01145	0.931	0.01775	0.922	0.08548	0.924
	AIC	0.01132	0.933	0.01728	0.897	0.08109	0.835
	BIC	0.01134	0.930	0.01841	0.869	0.08719	0.797
$n = 300$	SFIC	0.00779	0.934	0.01027	0.946	0.04253	0.943
	FIC	0.00790	0.933	0.01112	0.943	0.05159	0.943
	AIC	0.00778	0.932	0.01089	0.910	0.04958	0.855
	BIC	0.00787	0.934	0.01179	0.885	0.05421	0.807

相当, 但是 FIC 的经验覆盖概率明显比 AIC 和 BIC 高, 尤其是对 ϕ_3 而言.

(2) 在固定的删失比例下, 随着样本量的增加, SFIC, FIC, AIC 和 BIC 方法均越来越精确. 总体来说, 优劣性对比与 (1) 基本相同.

(3) 显然, 删失比例对 4 种方法的估计精度均有显著影响. 删失概率越小, 估计的结果越精确.

5 实际数据分析

本节用所提方法分析原发性胆汁性肝硬化数据集. 该数据集来自于 1974–1984 年 Mayo Clinic 进行的原发性胆汁性肝硬化试验. 该项试验的目的是研究原发性胆汁性肝硬化病人的存活时间与其影响因素之间的关系. 数据集包含了 312 位病人的存活时间和 17 个可能的影响因素, 在 17 个可能的影响因素中, 有 10 个连续型变量: 年龄 (age)、白蛋白含量 (alb)、碱性磷酸酶含量 (alk)、血清胆红素含量 (bil)、胆固醇含量 (chol)、尿铜 (cop)、血小板含量 (plat)、凝血酶原时间 (prot)、谷草转氨酶 (sgot)、甘油三酯 (trig) 和 7 个离散型变量: 是否存在腹水 (asc)、水肿处理情况 (ede)、是否有肝肿大 (hep)、性别 (sex)、是否有蛛网物 (spid)、疾病阶段 (stage)、控制组和处理组 (trt). 该数据集被许多学者研究过, 如文献 [4, 32, 35] 等. 在 312 个病人中, 有 36 个病人的部分影响因素数据出现了缺失, 剩余的 276 个病人的所有影响因素数据均被观测到. 与文献 [4, 32, 35] 一样, 我们在数据分析时仅利用了 276 个可完全观测个体的相关数据进行统计建模, 并且把病人的存活时间取了自然对数, 且标准化了连续性变量. 考虑到文献 [32] 曾经用稀疏删失中位数变量选择的方法选出了清胆红素含量 (bil)、碱性磷酸酶含量 (alk)、白蛋白含量 (alb)、尿铜 (cop) 和凝血酶原时间 (prot) 5 个解释变量, 我们在模型平均时把这 5 个变量设定为每个子模型的必选变量, 其余的 12 个变量用来做模型平均. 我们分别用 SFIC, FIC, AIC 和 BIC 方法计算了每个解释变量的回归系数及其置信水平为 95% 的置信区间, 结果见表 3. 从表 3 可以看出, SFIC 和 FIC 方法进一步肯定了白蛋白含量 (alb) 对病人存活时间的影响, 而 AIC 和 BIC 方法则低估了白蛋白含量 (alb) 对病人存活时间的影响作用; 同时, 与 SFIC 和 FIC 方法的结果相比, AIC 和 BIC 方法较大幅度地高估了是否存在腹水 (asc) 对病人寿命时间的影响作用.

表 3 原发性胆汁性肝硬化数据集的估计及置信区间

变量	AIC	BIC	FIC	SFIC
bil	-0.058 (-0.094 -0.022)	-0.063 (-0.098 -0.028)	-0.057 (-0.083 -0.030)	-0.059 (-0.084 -0.027)
alk	0.284 (0.147 0.421)	0.327 (0.186 0.468)	0.312 (0.178 0.446)	0.336 (0.173 0.449)
alb	0.534 (0.251 0.817)	0.610 (0.328 0.891)	0.623 (0.366 0.881)	0.656 (0.325 0.836)
cop	-0.275 (-0.393 -0.157)	-0.301 (-0.426 -0.176)	-0.241 (-0.329 -0.153)	-0.277 (-0.351 -0.162)
prot	0.136 (0.021 0.251)	0.153 (0.049 0.257)	0.123 (0.033 0.213)	0.119 (0.019 0.203)
ede	-0.495 (-1.508 0.518)	0	0	-0.191 (-0.749 -0.252)
plat	0	0	0	0.002 (-0.056 -0.015)
asc	-0.464 (-1.338 0.410)	-0.723 (-1.535 0.088)	0	-0.161 (-0.614 -0.236)
age	0.200 (-0.020 0.419)	0	0.238 (0.055 0.420)	0.185 (0.042 0.374)
trt	0	0	0	0.001 (-0.022 0.025)
chol	0.330 (0.099 0.561)	0.492 (0.178 0.807)	0.423 (0.204 0.641)	0.419 (0.163 0.551)
sex	0	0	0	0.023 (0.076 0.227)
sgot	0	0	0	0.006 (-0.049 0.002)
trig	0	0	0	-0.004 (-0.107 -0.048)
stage	-0.144 (-0.240 -0.048)	0	-0.130 (-0.213 -0.046)	-0.128 (-0.228 -0.054)
hep	0.172 (-0.052 0.396)	0	0.160 (-0.031 0.350)	0.125 (0.034 0.359)
spid	-0.168 (-0.395 0.059)	-0.260 (-0.510 -0.011)	-0.219 (-0.424 -0.015)	-0.175 (-0.394 0.005)

6 结论

本文讨论了响应变量随机右删失情况下线性模型的 FIC 模型选择和光滑 FIC 模型平均估计方法, 通过对最小二乘目标函数进行无偏调整, 构造了每个子模型下兴趣参数的估计, 并基于子模型估计的渐近均方误差的估计, 构造了每个子模型的 FIC 值, 以进行模型选择和构造光滑 FIC 模型平均估计的权函数, 然后给出了兴趣参数的光滑 FIC 模型平均估计的渐近分布. 模拟结果也显示所提方法具有较好的有限样本性质.

然而, 本文仅仅考虑了经典的线性模型, 一个仍然需要深刻考虑的问题是, 线性模型有着当线性假定与实际情况相符时, 估计的精度较高的优点, 同时也有着模型假定与实际情况有较大偏离时, 估计的偏差也有较大的缺点, 稳健性不够理想. 因此, 进一步讨论一些更具灵活性的复杂模型, 如半参数回归模型是具有意义的研究课题. 半参数模型既包含参数部分也包含非参数部分. 一般来说, 非参数部分估计的收敛速度要比参数部分估计的收敛速度慢, 其理论研究难度也更大. 因此, 如何在随机右删失数据下将线性模型的模型平均方法深入到半参数模型也是一个极具挑战性的问题. 我们拟在将来的研究中对这类问题进行讨论, 以期得到一些更为深入的结果.

7 定理的证明

引理 4 记 $A_n = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{1-G(Y_i)} \Pi_S X_i X_i^T \Pi_S^T$. 假设条件 (A₁)-(A₆) 成立, 有

$$A_n \xrightarrow{P} \Pi_S \Sigma \Pi_S^T,$$

其中, $\xrightarrow{P}$ 表示以概率收敛.

证明 经简单计算可得

$$\begin{aligned} A_n &= \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{1-G(Y_i)} \Pi_S X_i X_i^T \Pi_S^T + \frac{1}{n} \sum_{i=1}^n \delta_i \Pi_S X_i X_i^T \Pi_S^T \left(\frac{1}{1-\hat{G}(Y_i)} - \frac{1}{1-G(Y_i)} \right) \\ &= A_{n1} + A_{n2}. \end{aligned}$$

对于 A_{n1} , 有

$$\begin{aligned} A_{n1} &= E \left\{ \frac{1}{n} \sum_{i=1}^n E \left[\frac{\delta_i}{1-G(Y_i)} \Pi_S X_i X_i^T \Pi_S^T \mid T_i, X_i \right] \right\} \\ &= E \left\{ \frac{1}{n} \sum_{i=1}^n \Pi_S X_i X_i^T \Pi_S^T E \left[\frac{\delta_i}{1-G(Y_i)} \mid T_i, X_i \right] \right\} \\ &= E \left\{ \frac{1}{n} \sum_{i=1}^n \Pi_S X_i X_i^T \Pi_S^T \right\} \xrightarrow{P} \Pi_S \Sigma \Pi_S^T. \end{aligned} \quad (7.1)$$

对于 A_{n2} , 有

$$\begin{aligned} A_{n2} &= \frac{1}{n} \sum_{i=1}^n \delta_i \Pi_S X_i X_i^T \Pi_S^T \left(\frac{1}{1-\hat{G}(Y_i)} - \frac{1}{1-G(Y_i)} \right) \\ &= \frac{1}{n} \sum_{i=1}^n \delta_i \Pi_S X_i X_i^T \Pi_S^T \frac{(\hat{G}_n(Y_i) - G(Y_i))^2}{(1-G(Y_i))^3} \\ &\quad + \frac{1}{n} \sum_{i=1}^n \delta_i \Pi_S X_i X_i^T \Pi_S^T \frac{(\hat{G}_n(Y_i) - G(Y_i))^3}{(1-\hat{G}_n(Y_i))(1-G(Y_i))^3} \\ &\quad + \frac{1}{n} \sum_{i=1}^n \delta_i \Pi_S X_i X_i^T \Pi_S^T \frac{\hat{G}_n(Y_i) - G(Y_i)}{(1-G(Y_i))^2} \\ &\triangleq A_{n21} + A_{n22} + A_{n23}. \end{aligned}$$

注意到, Zhou^[36] 证明了

$$\sup_{0 \leq y \leq \hat{Y}_{(n)}} \left| \frac{\hat{G}_n(y) - G(y)}{1-G(y)} \right| = O_p(n^{-\frac{1}{2}}), \quad (7.2)$$

$$\sup_{0 \leq y \leq \hat{Y}_{(n)}} \left| \frac{\hat{G}_n(y) - G(y)}{1-\hat{G}_n(y)} \right| = O_p(1). \quad (7.3)$$

由 (7.2) 和 (7.3), 不难得到

$$A_{n2u} = o_p(1), \quad u = 1, 2, 3.$$

因此, 得

$$A_{n2} = o_p(1). \quad (7.4)$$

由 (7.1) 和 (7.4), 立得引理 4. □

定理 1 的证明 经简单计算, 不难得到

$$\sqrt{n}(\hat{\beta}_S - \Pi_S \beta_0) = A_n^{-1} \cdot \frac{1}{\sqrt{n}} \left[\sum_{i=1}^n \frac{\delta_i}{1-\hat{G}(Y_i)} \Pi_S X_i Y_i - \sum_{i=1}^n \frac{\delta_i}{1-\hat{G}(Y_i)} \Pi_S X_i X_i^T \Pi_S^T \Pi_S \beta_0 \right]$$

$$\begin{aligned}
&= A_n^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\delta_i}{1 - \hat{G}(Y_i)} \Pi_S X_i (Y_i - X_i^T \Pi_S^T \Pi_S \beta_0) \\
&= A_n^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\delta_i}{1 - \hat{G}(Y_i)} \Pi_S X_i (\varepsilon_i + X_i^T (I - \Pi_S^T \Pi_S) \beta_0) \\
&= A_n^{-1} \cdot (B_{n1} + B_{n2}),
\end{aligned}$$

其中,

$$\begin{aligned}
B_{n1} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\delta_i}{1 - \hat{G}(Y_i)} \Pi_S X_i \varepsilon_i, \\
B_{n2} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\delta_i}{1 - \hat{G}(Y_i)} \Pi_S X_i X_i^T (I - \Pi_S^T \Pi_S) \beta_0.
\end{aligned}$$

对于 B_{n1} , 有

$$\begin{aligned}
B_{n1} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\delta_i}{1 - G(Y_i)} \Pi_S X_i \varepsilon_i + \frac{1}{\sqrt{n}} \sum_{i=1}^n \delta_i \Pi_S X_i \varepsilon_i \left(\frac{1}{1 - \hat{G}(Y_i)} - \frac{1}{1 - G(Y_i)} \right) \\
&= B_{n11} + B_{n12}.
\end{aligned}$$

对于 B_{n12} , 有

$$\begin{aligned}
B_{n12} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \delta_i \Pi_S X_i \varepsilon_i \frac{(\hat{G}_n(Y_i) - G(Y_i))^2}{(1 - G(Y_i))^3} \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \delta_i \Pi_S X_i \varepsilon_i \frac{(\hat{G}_n(Y_i) - G(Y_i))^3}{(1 - \hat{G}_n(Y_i))(1 - G(Y_i))^3} + \frac{1}{\sqrt{n}} \sum_{i=1}^n \delta_i \Pi_S X_i \varepsilon_i \frac{\hat{G}_n(Y_i) - G(Y_i)}{(1 - G(Y_i))^2} \\
&\triangleq B_{n121} + B_{n122} + B_{n123}.
\end{aligned}$$

结合 (7.2) 和 (7.3) 易得

$$B_{n121} = o_p(1), \quad B_{n122} = o_p(1). \quad (7.5)$$

对于 B_{n123} , 有

$$\begin{aligned}
B_{n123} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\delta_i \Pi_S X_i \varepsilon_i}{1 - G(Y_i)} \frac{\hat{G}_n(Y_i) - G(Y_i)}{1 - G(Y_i)} + o_p(1) \\
&= \frac{1}{n} \sum_{i=1}^n \frac{\delta_i \Pi_S X_i \varepsilon_i}{1 - G(Y_i)} \frac{1}{\sqrt{n}} \sum_{j=1}^n \int_0^{\tau_Q} \frac{I[Y_i > y]}{(1 - G(y))(1 - F(y))} dA_j(y) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^{\tau_Q} \frac{E[\Pi_S X \varepsilon I[Y > y]]}{(1 - G(y))(1 - F(y))} dA_i(y) + o_p(1),
\end{aligned}$$

其中,

$$A_i(y) = I[Y_i \leq y, \delta_i = 0] - \int_0^y I[Y_i \geq s] d\Lambda_G(s),$$

且 $\{A_i(y), 0 < y < \infty\}$ 为鞅, 其可料变差过程为 $\langle A_i \rangle(y) = \int_{-\infty}^y I[Y_i \geq s] d\Lambda_G(s)$. 因此,

$$B_{n12} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^{\tau_Q} \frac{E[\Pi_S X \varepsilon I[Y > y]]}{(1 - G(y))(1 - F(y))} dA_i(y) + o_p(1). \quad (7.6)$$

结合 (7.5) 和 (7.6), 有

$$\begin{aligned} B_{n11} + B_{n12} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\delta_i}{1 - G(Y_i)} \Pi_S X_i \varepsilon_i + \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^{\tau_Q} \frac{E[\Pi_S X \varepsilon I[Y > y]]}{(1 - G(y))(1 - F(y))} dA_i(y) + o_p(1) \\ &\triangleq \Pi_S(R_{n1} + R_{n2}) + o_p(1). \end{aligned}$$

易见, 对 R_{n1} 的协方差阵, 有

$$\text{Var}(R_{n1}) = E \left[\frac{X X^T \varepsilon^2}{1 - G(Y)} \right].$$

下面考虑 R_{n2} . 记

$$R_{n2}(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^u \frac{E[X \varepsilon I[Y > y]]}{(1 - G(y))(1 - F(y))} dA_i(y), \quad 0 \leq u \leq \tau_Q,$$

则 $R_{n2} = R_{n2}(\tau_Q)$. 并且, $\{R_{n2}(u), 0 < u \leq \tau_Q\}$ 为一个局部鞅, 其可料变差过程为

$$\langle R_{n2}(u) \rangle = \int_0^u (B(y))^{\otimes 2} \left(\frac{1}{n} \sum_{i=1}^n I[Y_i \geq y] \right) d\Lambda_G(y) \rightarrow \int_0^u (B(y))^{\otimes 2} (1 - F(y)) dG(y) \triangleq \Omega_2(u).$$

注意到对于充分大的 n , 有 $I[\frac{1}{\sqrt{n}} \|B(y)\| \geq \epsilon] \stackrel{\text{a.s.}}{\rightarrow} 0$, 易得如下的 Lindeberg 条件成立, 对于任意的 ϵ , 有

$$\sum_{i=1}^n \int_0^u \left(\frac{1}{\sqrt{n}} B(y) \right)^{\otimes 2} I \left[\frac{1}{\sqrt{n}} \|B(y)\| \geq \epsilon \right] I[Y_i \geq y] d\Lambda_G(y) \xrightarrow{P} 0.$$

因此, 由 Rebolledo 鞅的中心极限定理^[37] 可知, $\{R_{n2}(u), 0 < u \leq \tau_Q\}$ 弱收敛到一个 Gauss 鞅. 类似于文献 [38, (2.29)], 可证

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \int_u^{\tau_Q} \frac{E[X \varepsilon I[Y > y]]}{(1 - G(y))(1 - F(y))} dA_i(y) \xrightarrow{P} 0.$$

因此, $\text{Var}(R_{n2}) \xrightarrow{P} \Omega_2(\tau_Q)$. 为了得到 $R_{n2} + R_{n1}$ 的渐近分布, 考虑向量 $(R_{n1}^T, R_{n2}^T(u))^T$. 注意到

$$(R_{n1}^T, R_{n2}^T(u))^T = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\phi_i^T, \left(\int_0^u \frac{E[X \varepsilon I[Y > y]]}{(1 - G(y))(1 - F(y))} dA_i(y) \right)^T \right)^T \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^n (\phi_i^T, \varphi_i^T)^T$$

为独立 0 均值随机向量的和. 由 (A₅) 和 $\int_0^u B(y) dA_i(y)$ 的一致有界性, 可得

$$I \left[\frac{1}{n} \|(\phi_i^T, \varphi_i^T)^T\|^2 \geq \epsilon \right] \stackrel{\text{a.s.}}{\rightarrow} 0.$$

因此, Lindeberg 条件

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n E \|(\phi_i^T, \varphi_i^T)^T\|^2 \cdot I \left[\frac{1}{n} \|(\phi_i^T, \varphi_i^T)^T\|^2 \geq \epsilon \right] = 0$$

成立. 由独立向量和的中心极限定理知, $(R_{n1}^T, R_{n2}^T(u))^T$ 的极限分布为正态分布. 令 $u \rightarrow \tau_Q$ 可知, $R_{n1} + R_{n2}(\tau_Q)$ 的极限也是正态分布. 下面考虑 R_{n2} 和 R_{n1} 的协方差阵. 注意到 $\|\sum_{i=1}^n \int_0^u B(y) dA_i(y)\|$ 为一致有界的, 且

$$E \left[\int_0^u B(y) dA_i(y) \right] = 0, \quad \delta_i(1 - \delta_i) = 0, \quad dA_i(y) = (1 - \delta_i) dI[Y_i \leq y] - I[Y_i \geq y] d\Lambda_G(y),$$

可得

$$\begin{aligned}
 E[R_{n1}R_{n2}^T(u)] &= \frac{1}{n}E\left\{\sum_{i=1}^n \frac{\delta_i}{1-G(Y_i)} X_i \varepsilon_i \int_0^u B^T(y) dA_i(y)\right\} \\
 &= \int_0^u E\left\{\frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{1-G(Y_i)} X_i \varepsilon_i B^T(y) (1-\delta_i) dI[Y_i \leq y]\right\} \\
 &\quad - \int_0^u E\left\{\frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{1-G(Y_i)} X_i \varepsilon_i B^T(y) I[Y_i \geq y]\right\} d\Lambda_G(y) \\
 &= - \int_0^u E\left\{\frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{1-G(Y_i)} X_i \varepsilon_i B^T(y) I[Y_i \geq y]\right\} d\Lambda_G(y) \\
 &\rightarrow - \int_0^u E\{X_i \varepsilon_i B^T(y) I[Y_i \geq y]\} d\Lambda_G(y) \\
 &\triangleq \Omega_3(u).
 \end{aligned}$$

至此, 有

$$B_{n1} = B_{n11} + B_{n12} + o_p(1) = \Pi_S W + o_p(1),$$

其中, $W \sim N_d(0, \Omega)$, $\Omega = \Omega_1 + \Omega_2 + 2\Omega_3$.

下面考虑 B_{n2} . 由 $\hat{G}$ 的一致相合性, 类似于引理 4, 且注意到

$$(I - \Pi_S^T \Pi_S) \begin{pmatrix} \ddot{\beta}_0 \\ 0 \end{pmatrix} = 0,$$

容易证明

$$\begin{aligned}
 B_{n2} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\delta_i}{1-G(Y_i)} \Pi_S X_i X_i^T (I - \Pi_S^T \Pi_S) \beta_0 \\
 &\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\delta_i \Pi_S X_i X_i^T}{1-G(Y_i)} (I - \Pi_S^T \Pi_S) \beta_0 \frac{1}{n} \sum_{j=1}^n \int_0^{\tau_Q} \frac{I[Y_j > y]}{(1-G(y))(1-F(y))} dA_j(y) + o_p(1) \\
 &= \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{1-G(Y_i)} \Pi_S X_i X_i^T (I - \Pi_S^T \Pi_S) \sqrt{n} \beta_0 + o_p(1) \\
 &\quad + \frac{1}{\sqrt{n}} \cdot \frac{1}{n} \sum_{i=1}^n \frac{\delta_i \Pi_S X_i X_i^T}{1-G(Y_i)} (I - \Pi_S^T \Pi_S) \begin{pmatrix} 0 \\ \eta \end{pmatrix} \\
 &\quad \times \frac{1}{\sqrt{n}} \sum_{j=1}^n \int_0^{\tau_Q} \frac{I[Y_j > y]}{(1-G(y))(1-F(y))} dA_j(y) + o_p(1) \\
 &= \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{1-G(Y_i)} \Pi_S X_i X_i^T (I - \Pi_S^T \Pi_S) \begin{pmatrix} 0 \\ \eta \end{pmatrix} + \frac{1}{\sqrt{n}} \cdot O_p(1) + o_p(1) \\
 &= \Pi_S \Sigma (I - \Pi_S^T \Pi_S) \begin{pmatrix} 0 \\ \eta \end{pmatrix} + o_p(1).
 \end{aligned}$$

结合引理 4 和 slusky 定理, 不难得到

$$\sqrt{n}(\hat{\beta}_S - \Pi_S \beta_0) \xrightarrow{d} (\Pi_S \Sigma \Pi_S^T)^{-1} \Pi_S W + (\Pi_S \Sigma \Pi_S^T)^{-1} \Pi_S \Sigma (I - \Pi_S^T \Pi_S) \begin{pmatrix} 0 \\ \eta \end{pmatrix}. \quad (7.7)$$

对 (7.7) 进行简单变形, 不难得到定理 1. 证毕. $\square$

定理 2 的证明 基于定理 1, 由 δ 方法和文献 [39, 定理 3.2] 直接得到定理 2. $\square$

定理 3 的证明

$$\begin{aligned} \sqrt{n}(\hat{\phi} - \phi_0) &= \sum_S w(S | \hat{\eta}) \sqrt{n}(\hat{\phi}_S - \phi_0) \\ &= \sum_S w(S | \hat{\eta}) \left\{ \phi'_{\beta_0} \Pi_S^T (\Pi_S \Sigma \Pi_S^T)^{-1} \Pi_S W + \phi'_{\beta_0} A_S \begin{pmatrix} 0 \\ \eta \end{pmatrix} + o_p(1) \right\} \\ &= \sum_S w(S | \hat{\eta}) \phi'_{\beta_0} \Pi_S^T (\Pi_S \Sigma \Pi_S^T)^{-1} \Pi_S W + \sum_S w(S | \hat{\eta}) \phi'_{\beta_0} A_S \begin{pmatrix} 0 \\ \eta \end{pmatrix} + o_p(1). \end{aligned}$$

注意到 $\hat{\eta} \xrightarrow{d} \Delta$, 由连续映射定理知, $w(S | \hat{\eta}) \xrightarrow{d} w(S | \Delta)$. 由此立得定理 3. 证毕. $\square$

致谢 感谢审稿人的细致阅读以及他们深刻的建议和指导.

参考文献

- 1 Akaike H. Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, 1973, 60: 255–265
- 2 Schwarz G. Estimating the dimension of a model. *Ann Statist*, 1978, 6: 461–464
- 3 Mallows C L. Some comments on C_p . *Technometrics*, 1973, 15: 661–675
- 4 Tibshirani R. Regression shrinkage and selection via the lasso. *J R Statist Soc B*, 1996, 58: 267–288
- 5 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Statist Assoc*, 2001, 96: 1348–1360
- 6 Claeskens G, Hjort N L. The focused information criterion. *J Amer Statist Assoc*, 2003, 98: 900–916
- 7 Claeskens G, Croux C, Van Kerckhoven J. Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics*, 2006, 62: 972–979
- 8 Hjort N L, Claeskens G. Focussed information criteria and model averaging for Cox's hazard regression model. *J Amer Statist Assoc*, 2006, 101: 1449–1464
- 9 Zhang X Y, Liang H. Focused information criterion and model averaging for generalized additive partial linear models. *Ann Statist*, 2011, 39: 174–200
- 10 Du J, Zhang Z Z, Xie T F. Model averaging in quantile regression. *Comm Statist Theory Methods*, in press, 2011
- 11 Yuan Z, Yang Y. Combining linear regression models: when and how? *J Amer Statist Assoc*, 2005, 4: 1202–1214
- 12 Leeb H, Pötscher B M. The finite sample distribution of post-model-selection estimators and uniform versus non-uniform approximations. *Econometric Theory*, 2003, 19: 100–142
- 13 Leeb H, Pötscher B M. Can one estimate the unconditional distribution of post-model-selection estimators. *Econometric Theory*, 2008, 24: 38–376
- 14 Leung G, Barron A R. Information theory and mixing least- squares regressions. *IEEE Trans Inform Theory*, 2006, 52: 3396–3410
- 15 Yang, Y. Adaptive regression by mixing. *J Amer Statist Assoc*, 2001, 96: 574–586
- 16 Bates J, Granger C. The combination of forecasts. *Oper Res Soc*, 1969, 20: 451–468
- 17 Buckland S T, Burnham K P, Augustin N H. Model selection: an integral part of inference. *Biometrics*, 1997, 53: 603–618

- 18 Hansen B. Least squares model averaging. *Econometrica*, 2007, 4: 1175–1189
- 19 Hjort N L, Claeskens G. Frequentist model average estimators. *J Amer Statist Assoc*, 2003, 98: 879–899
- 20 Liang H, Zou G H, Wan A T K, et al. On optimal weight choice in a frequentist model average estimator. *J Amer Statist Assoc*, 2011, 106: 1053–1066
- 21 Deng G H, Liang H. Model averaging for semiparametric additive partial linear models. *Sci China Math*, 2010, 53: 1363–1376
- 22 Sun Z M. Factor model averaging quantile regression and simulation study. In: *Lecture Notes in Comput Sci*, vol. 7473. Berlin: Springer, 2012, 291–298
- 23 Buckley J. Linear regression with censored data. *Biometrika*, 1979, 66: 429–436
- 24 Koul H, Susarla V V, Ryzin J. Regression analysis with randomly right-censored data. *Ann Statist*, 1981, 9: 1276–1288
- 25 Leurgans S. Linear models, random censoring and synthetic data. *Biometrika*, 1987, 74: 301–309
- 26 Wang Q H, Li G. Empirical likelihood semiparametric regression analysis under random censorship. *J Multivariate Anal*, 2002, 83: 469–486
- 27 Stute W. Consistent estimation under random censorship when covariables are present. *J Multivariate Anal*, 1993, 45: 89–103
- 28 Stute W. Distributional convergence under random censorship when covariables are present. *Scand J Statist*, 1996, 23: 461–471
- 29 Li X Y, Wang Q H. The weighted least square based estimators with censoring indicators missing at random. *J Statist Plann Inference*, 2012, 142: 2913–2925
- 30 Bang H, Tsiatis A. Median regression with censored cost data. *Biometrics*, 2002, 58: 643–649
- 31 Peng L M, Huang Y J. Survival analysis with quantile regression models. *J Amer Statist Assoc*, 2008, 103: 637–649
- 32 Shows J H, Lu W B, Zhang H H. Sparse estimation and inference for censored median regression. *J Statist Plann Inference*, 2010, 140: 1903–1917
- 33 Sun Z M, Zhang Z Z, Du J. Semiparametric analysis of isotonic errors-in-variables regression models with randomly right censored response. *J Syst Sci Complex*, 2013, 26: 1–21
- 34 Kaplan E L, Meier P. Nonparametric estimation from incomplete observations. *J Amer Statist Assoc*, 1958, 53: 457–481
- 35 Zhang H, Lu W. Adaptive LASSO for Cox’s proportional hazards model. *Biometrika*, 2007, 94: 1–13
- 36 Zhou M. Asymptotic normality of the synthetic estimator for censored survival data. *Ann Statist*, 1992, 20: 1002–1021
- 37 Shorack G R, Wellner J A. *Empirical Processes with Applications to Statistics*. New York: John Wiley and Sons, 1986
- 38 Lai T L, Ying Z L, Zheng Z K. Asymptotic normality of a class of adaptive statistics with applications to synthetic data method for censored regression. *J Multivariate Anal*, 1995, 52: 259–279
- 39 van der Vaart A W, Wellner J A. *Weak Convergence and Empirical Processes: with Application to Statistics*. New York: Springer-Verlag, 1996

FIC model selection and model averaging for linear model with censored response

SUN ZhiMeng, MA JingYi & SU Zhi

Abstract In this paper, we consider the FIC model selection and model averaging approach for the censored linear model. We show the asymptotic normality of the proposed FIC model selection and smoothed FIC model averaging estimator of the focused parameter. We carry a simulation study to investigate the finite property of the proposed estimator. It results out that the proposed estimator is superior to some traditional model selection methods, like AIC and BIC, in light of MSE of the estimator and the empirical coverage probability of confidence interval. We illustrated the proposed method via an analysis of a primary biliary cirrhosis (BPC) data.

Keywords randomly right censored, model selection, model averaging, estimation theory

MSC(2010) 62N01, 62G05

doi: 10.1360/012012-456